GETTING STARTED

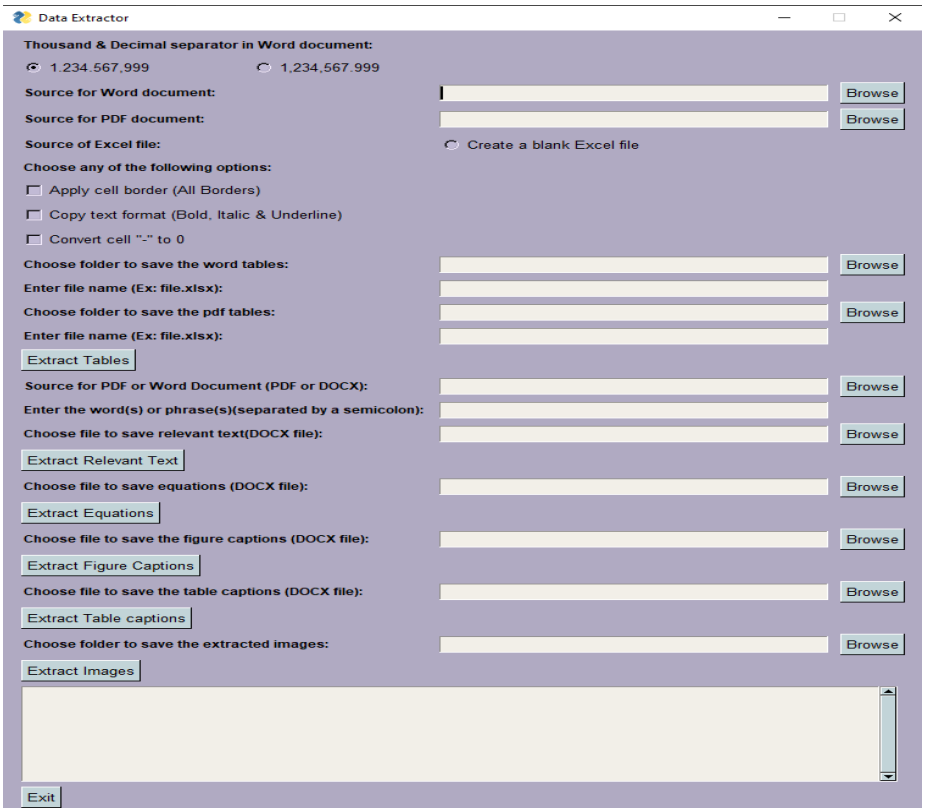
This is a data extraction tool with a Graphical User Interface (GUI). It provides various functionalities to extract data from PDF and Word documents, including tables, equations, relevant text, figure captions, and images.

➢ Necessary libraries to be installed (pip install requirement.txt)

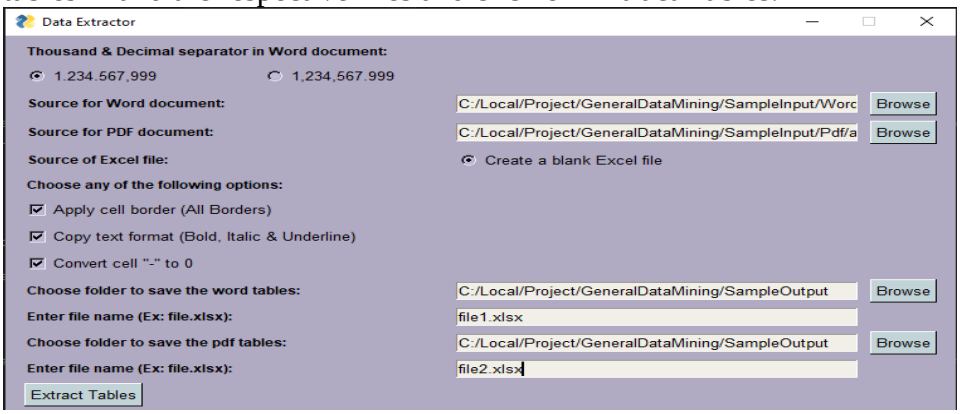
1) PySimpleGUI: A Python GUI framework for creating simple and easy-to-use graphical user interfaces.
   You can install it using the following command: **pip install PySimpleGUI**
2) PyPDF2: A library for working with PDF files in Python. You can install PyPDF2 using the following command: **pip install PyPDF2**
3) docx2txt: A library for extracting text from Microsoft Word (DOCX) files.
   You can install it using the following command: **pip install docx2txt**
4) locale: A module to set and access locale-specific information in Python. The locale module is a built-in module in Python and should be available by default.
5) openpyxl: A library for working with Excel files in Python. You can install it using the following command: **pip install openpyxl**
6) tabula-py: A library for extracting tables from PDF documents in Python. You can install it using the following command: **pip install tabula-py**
7) pandas: A powerful library for data manipulation and analysis in Python. You can install it using the following command: **pip install pandas**
8) fitz: A library for working with PDF files in Python, part of the PyMuPDF package. You can install it using the following command: **pip install PyMuPDF**
9) Pillow: A library for working with images in Python. You can install it using the following command: **pip install Pillow**

➢ How to run the code

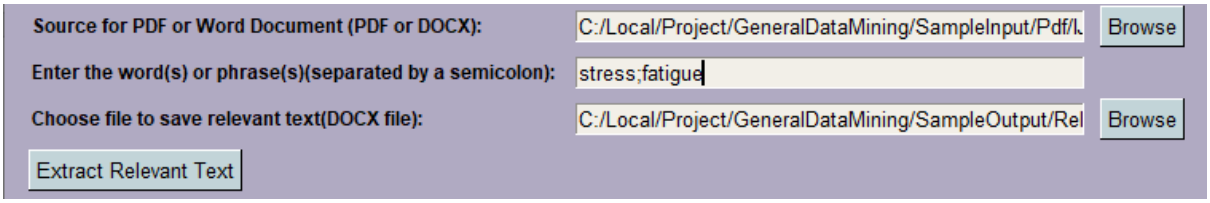
1) Clone or download this repository to your local machine.
2) Open a terminal or command prompt and navigate to the directory containing the script and input files (if applicable).
3) Run the script using Python, and it will ask you the following.



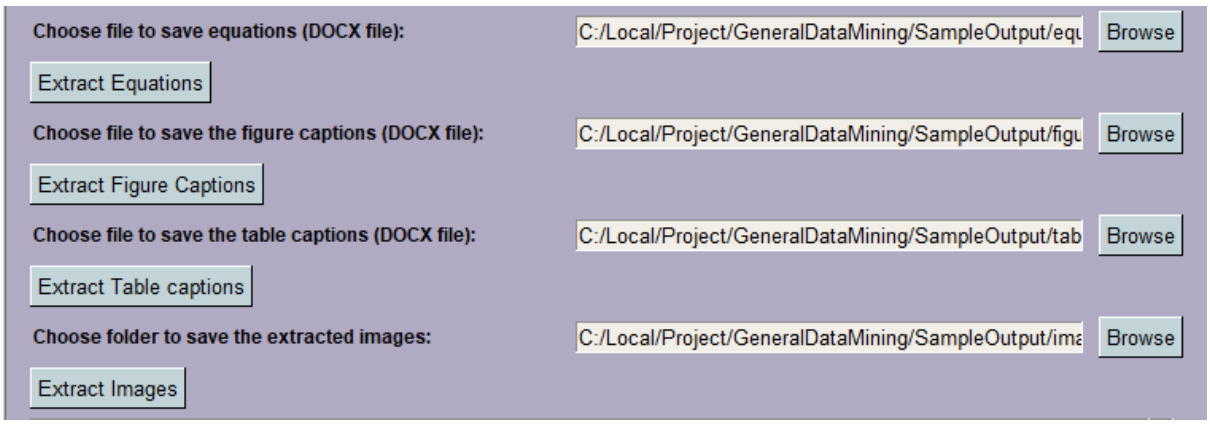
4) Select the source of word document and pdf document, then choose the below options such as apply cell border etc. Then choose the folder you want to save the extracted tables in and the respective files and click on Extract Tables.



5) To extract the relevant text related to words or phrases type the words separated by a semicolon and click extract relevant text.



6) To extract Equations, figure captions, table captions and images do as shown below.



7) Once Execution is finished the following will be displayed then click on exit.

➢ Code Explanation

GUI Layout: The script defines the layout for the GUI using the PySimpleGUI library. The layout includes several input fields, checkboxes, and buttons to facilitate user interactions.

Data Extraction Functions: The script includes multiple functions for extracting different types of data from PDF and Word documents.

Linking GUI to Data Extraction: The script connects the GUI buttons with their respective data extraction functions. When a user clicks a button, the corresponding data extraction function is called with the provided input values from the GUI.

Main Loop: The script enters a main loop using while True to handle events from the GUI. It listens for user actions, such as clicking buttons or closing the GUI window, and executes the corresponding data extraction functions accordingly.

Output and Exit: After each data extraction process, the script prints messages to the console indicating the completion of the specific task. The program will exit either when the user closes the GUI window or clicks the 'Exit' button.

Overall, this script provides a user-friendly interface to extract specific types of data from PDF and Word documents, helping users to work with and analyze data more effectively. Users can choose various options, such as specifying the source files, output formats, and other preferences, using the GUI provided by the PySimpleGUI library.